openai.com

How should AI systems behave, and who should decide?

Authors

3-4 minutes

In pursuit of our mission, we're committed to ensuring that access to, benefits from, and influence over AI and AGI are widespread. We believe there are at least three building blocks required in order to achieve these goals in the context of AI system behavior.[B]

1. Improve default behavior. We want as many users as possible to find our AI systems useful to them "out of the box" and to feel that our technology understands and respects their values.

Towards that end, we are investing in research and engineering to reduce both glaring and subtle biases in how ChatGPT responds to different inputs. In some cases ChatGPT currently refuses outputs that it shouldn't, and in some cases, it doesn't refuse when it should. We believe that improvement in both respects is possible.

Additionally, we have room for improvement in other dimensions of system behavior such as the system "making things up." Feedback from users is invaluable for making these improvements.

2. Define your AI's values, within broad bounds. We believe that AI should be a useful tool for individual people, and thus

customizable by each user up to limits defined by society. Therefore, we are developing an upgrade to ChatGPT to allow users to easily customize its behavior.

This will mean allowing system outputs that other people (ourselves included) may strongly disagree with. Striking the right balance here will be challenging–taking customization to the extreme would risk enabling malicious uses of our technology and sycophantic AIs that mindlessly amplify people's existing beliefs.

There will therefore always be some bounds on system behavior. The challenge is defining what those bounds are. If we try to make all of these determinations on our own, or if we try to develop a single, monolithic AI system, we will be failing in the commitment we make in our Charter to "avoid undue concentration of power."

3. Public input on defaults and hard bounds. One way to avoid undue concentration of power is to give people who use or are affected by systems like ChatGPT the ability to influence those systems' rules.

We believe that many decisions about our defaults and hard bounds should be made collectively, and while practical implementation is a challenge, we aim to include as many perspectives as possible. As a starting point, we've sought external input on our technology in the form of red teaming. We also recently began soliciting public input on AI in education (one particularly important context in which our technology is being deployed).

We are in the early stages of piloting efforts to solicit public input on topics like system behavior, disclosure mechanisms (such as watermarking), and our deployment policies more broadly. We are

also exploring partnerships with external organizations to conduct third-party audits of our safety and policy efforts.